

# On Multilinear Principal Component Analysis of Order-Two Tensors

Hung Hung<sup>a</sup>, Pei-Shien Wu<sup>b</sup>, I-Ping Tu<sup>c,\*</sup>, and Su-Yun Huang<sup>c</sup>

<sup>a</sup>Institute of Epidemiology and Preventive Medicine, National Taiwan University

<sup>b</sup>Department of Mathematics, National Taiwan University

<sup>c</sup>Institute of Statistical Science, Academia Sinica

## Abstract

Principal Component Analysis (PCA) is a commonly used tool for dimension reduction in analyzing high dimensional data; Multilinear Principal Component Analysis (MPCA) has the potential to serve the similar function for analyzing tensor structure data. MPCA and other tensor decomposition methods have been proved effective to reduce the dimensions for both real data analyses and simulation studies (Ye, 2005; Lu, Plataniotis and Venetsanopoulos, 2008; Kolda and Bader, 2009; Li, Kim and Altman, 2010). In this paper, we investigate MPCA's statistical properties and provide explanations for its advantages. Conventional PCA, vectorizing the tensor data, may lead to inefficient and unstable prediction due to its extremely large dimensionality. On the other hand, MPCA, trying to preserve the data structure, searches for low-dimensional multilinear projections and decreases the dimensionality efficiently. The asymptotic theories for order-two MPCA, including asymptotic distributions for principal components, associated projections and the explained variance, are developed. Finally, MPCA is shown to improve conventional PCA on analyzing the Olivetti Faces data set, by constructing more module oriented basis in reconstructing the test faces.

---

\*Corresponding author. *Email address:* iping@stat.sinica.edu.tw

**Keywords and phrases:** Asymptotic theory, Dimension reduction, Image reconstruction, Multilinear principal component analysis, Principal component analysis, Tensor.

# 1 Introduction

Dimension reduction is a key step for high dimensional data analysis. Principal component analysis (PCA) is probably the most commonly used method for dimension reduction. Given  $n$  observations on  $m$  variables, PCA calculates the  $m \times m$  covariance matrix and solves the eigenvalue decomposition problem for the covariance matrix. The goal is to choose a smaller set of eigenvectors as a new coordinate system so that the newly transformed variables can retain the most data variation. This PCA approach has been widely applied in many scientific fields for dimension reduction and compact data representation (Jolliffe, 2002), where the collected data are organized in an  $n \times m$  design matrix with each row representing an observation and each column a variable.

When data are tensor objects, traditional analysis vectorizes each of the tensor objects into a long vector and arranges these vectorized objects in a design matrix form. Subsequent analysis is followed in the usual way. Nevertheless, this approach usually produces a large number of variables, where the available sample size is relatively small, and many existing statistical methods fail to apply. For a typical example, like the Olivetti Faces data set to be used in an experimental study later, there are 400 images each with  $64 \times 64$  pixels. Vectorizing each image leads to a design matrix of size  $400 \times 4096$ , which the variable dimension  $m$  largely exceeds the sample size  $n$ .

One strategy to overcome this difficulty is to take advantage of the natural tensor structure of the data. Singular value decomposition (SVD) is an example. Given a  $p \times q$  matrix  $X$  which can be treated as an order-two tensor, SVD can decompose two directional spaces simultaneously:  $X = USV^T = \sum_{i=1}^{p \wedge q} s_i u_i v_i^T$ , where  $U = [u_1, \dots, u_p] \in \Re^{p \times p}$  and  $V = [v_1, \dots, v_q] \in \Re^{q \times q}$  are, respectively, the left and right singular vectors,  $S$  is a diagonal matrix of size  $p \times q$  with diagonal elements  $\{s_1, \dots, s_{(p \wedge q)}\}$ . The dimension can be reduced when the index  $i$  is properly truncated. De Lathauwer, De Moor and Vandewalle (2000a) then generalized the SVD to high-order SVD (HOSVD) for a given  $N^{\text{th}}$ -order tensor

object  $A \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ . Further, they formulated the problem of “best rank- $(R_1, \dots, R_N)$  approximation of higher-order tensors” in the least-squares sense, and discussed many algorithms to achieve this task (De Lathauwer, De Moor and Vandewalle, 2000b).

Later, Yang et al. (2004) proposed two-dimensional PCA (2DPCA) for analyzing image data, which are order-two tensors. An improved two-directional two-dimensional PCA  $((2D)^2PCA)$  was developed in Zhang and Zhou (2005), which was shown to perform better than 2DPCA through simulation studies. Ye (2005) formulated the problem of generalized low rank approximation of matrices, which can be treated as a sample extension of the best rank- $(R_1, R_2)$  approximation for order-two tensors in De Lathauwer, De Moor and Vandewalle (2000b). Lu, Plataniotis and Venetsanopoulos (2008) further generalized the work of Ye (2005) and proposed multilinear PCA (MPCA) for tensor objects of arbitrary orders. There are other tensor decomposition methods for dimension reduction. For instance, Kolda and Bader (2009) provided a general overview of current development of tensor decomposition methods for unsupervised learning, their applications, and available softwares; Li, Kim and Altman. (2010) considered the tensor decomposition methods for supervised learning such as regression and classification.

Similar to conventional PCA, the goal of MPCA is to look for low-dimensional multilinear projection for tensor objects that captures the most data variation. Back to the example of *Olivetti Faces*, one eigenvector in conventional PCA creates an image basis element that contains 4095 free parameters. By contrast, one image basis element in MPCA or  $(2D)^2PCA$ , which involves the Kronecker product of a column vector and a row vector, contains 126 free parameters. From the viewpoint of the number of parameters required to specify one basis element, MPCA is expected to perform better than conventional PCA, when the sample size is small to moderate, like this *Olivetti Faces* example. Compared to  $(2D)^2PCA$ , MPCA has the advantage of capturing more data variation by the chosen image basis, because of its specific criterion. MPCA has been successfully applied in real data analysis and checked by simulations (Ye, 2005; Lu et al., 2008). Yet, to our best knowledge, there is neither statistical justification nor asymptotic study for MPCA.

In this paper, we try to establish some relevant properties of order-two MPCA from a

statistical point of view. Our study is based on the following model:

$$X = \mu + A_0 U B_0^T + \varepsilon, \quad (1)$$

where  $\mu \in \mathbb{R}^{p \times q}$  is the mean parameter of  $X$ ,  $A_0 \in \mathbb{R}^{p \times p_0}$  and  $B_0 \in \mathbb{R}^{q \times q_0}$  with  $p_0 \leq p$  and  $q_0 \leq q$  are non-random basis matrices,  $U \in \mathbb{R}^{p_0 \times q_0}$  is a random coordinate matrix with  $E[U] = 0$  and a strictly positive definite covariance matrix  $\text{Cov}(\text{vec}(U)) = T \in \mathbb{R}^{m_0 \times m_0}$ , where  $m_0 = p_0 q_0$  and  $\text{vec}(\cdot)$  is the operator that stacks the columns of a tensor into a long vector. The error term  $\varepsilon \in \mathbb{R}^{p \times q}$  is a random matrix independent of  $U$  and with  $E[\varepsilon] = 0$  and  $\text{Cov}(\text{vec}(\varepsilon)) = \sigma^2 I_m$ , where  $m = pq$ . Under model (1) which characterizes the tensor structure of  $X$ , we justify the validity of MPCA. Asymptotic properties of MPCA are rigorously developed, including asymptotic distributions for principal components, associated projections and the explained variance. It is also shown that MPCA is asymptotically more efficient than (2D)<sup>2</sup>PCA in estimating the target dimension reduction subspace. Furthermore, a test of dimensionality is developed, based on the derived asymptotic results.

This paper is organized as follows. Section 2 presents some properties of the estimation for the target subspace and a test for its dimensionality. The relations between MPCA and both conventional PCA and (2D)<sup>2</sup>PCA are also discussed in this section. In section 3, the asymptotic theory of MPCA is developed. In section 4, the performance of MPCA and its comparison with conventional PCA is demonstrated by analyzing the Olivetti Faces data set. The paper ends with a brief discussion. Technical proofs of main results are deferred to the Appendix.

## 2 MPCA

MPCA, as a dimension reduction algorithm, is originally designed to search basis matrices  $\{A, B\}$  and coordinate matrices  $U_i$ 's that best approximate the observed data  $X_i$  as  $AU_iB^T$  for  $i = 1, \dots, n$ . Although many simulation studies and real data analyses in literature support the usage of MPCA and multilinear tensor decomposition (Ye, 2005; Lu et al., 2008; Kolda and Bader, 2009; Li et al., 2010), there is no theoretical study from the

statistical point of view. Let  $\otimes$  be the Kronecker product. Then, there is an equivalent formula for model (1)

$$X = \mu + A_0 U B_0^T + \varepsilon \Leftrightarrow \text{vec}(X - \mu) = (B_0 \otimes A_0) \text{vec}(U) + \text{vec}(\varepsilon) \quad (2)$$

by the fact that  $\text{vec}(A_0 U B_0^T) = (B_0 \otimes A_0) \text{vec}(U)$ . Without loss of generality, we may assume that  $A_0$  and  $B_0$  are orthogonal matrices, i.e.,  $A_0^T A_0 = I_{p_0}$  and  $B_0^T B_0 = I_{q_0}$ . Model (1) thus ensures that, without considering the error term  $\varepsilon$ , the columns and rows of  $(X - \mu)$  belong to  $\text{span}(A_0)$  and  $\text{span}(B_0)$ , respectively, and  $\text{vec}(X - \mu)$  belongs to the subspace  $\text{span}(B_0) \otimes \text{span}(A_0) = \text{span}(B_0 \otimes A_0)$ . It is then reasonable to estimate  $\text{span}(B_0 \otimes A_0)$  for follow-up analysis such as data compression, pattern recognition, regression analysis, etc. In this section, we show that, under model (1), MPCA actually attempts to extract a basis pair  $\{A, B\}$  targeting the subspace  $\text{span}(B_0 \otimes A_0)$ . Proposition 2.2 below proves the existence of a solution pair  $\{A, B\}$ . Proposition 2.5 summarizes that the inclusion relation between  $\text{span}(A)$  (resp.,  $\text{span}(B)$ ) and the target dimension reduction subspace  $\text{span}(A_0)$  (resp.,  $\text{span}(B_0)$ ), depends on the size comparison between the specified dimensionality  $\tilde{p}$  (resp.,  $\tilde{q}$ ) and  $p_0$  (resp.,  $q_0$ ). Recognizing the important roles of  $\tilde{p}$  and  $\tilde{q}$ , we construct a hypothesis test for choosing  $\tilde{p}$  and  $\tilde{q}$ . These works justify the usage of MPCA in extracting the relevant basis for subsequent analysis, provided the data has a natural tensor structure.

## 2.1 Estimation

Let  $\{X_i\}_{i=1}^n$  be the collected data set which are assumed to be random copies of a random matrix  $X \in \mathbb{R}^{p \times q}$ . MPCA aims to extract the basis pair that best approximate  $\{X_i\}_{i=1}^n$  while preserving the tensor structure of them. In particular, for a pre-specified dimensionality  $(\tilde{p}, \tilde{q})$ , Ye (2005) proposed a criterion to find  $A \in \mathcal{O}_{p, \tilde{p}}, B \in \mathcal{O}_{q, \tilde{q}}$ , and  $\{U_i\}_{i=1}^n$  that minimize

$$\frac{1}{n} \sum_{i=1}^n \|(X_i - \bar{X}) - A U_i B^T\|_F^2, \quad (3)$$

where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is the sample mean matrix,  $\|\cdot\|_F$  is the Frobenius norm of a matrix, and  $\mathcal{O}_{\ell, \tilde{\ell}}$  is the collection of all orthogonal matrices  $M$  of size  $\ell \times \tilde{\ell}$  such that  $M^T M = I_{\tilde{\ell}}$ .

Note that the objective function (3) can be expressed as

$$\frac{1}{n} \sum_{i=1}^n \|\text{vec}(X_i - \bar{X}) - (B \otimes A) \text{vec}(U_i)\|_2^2. \quad (4)$$

If we replace  $(B \otimes A)$  by  $\Gamma$  in (4), the minimization problem then becomes the conventional PCA. From this viewpoint, MPCA can be treated as a constrained PCA with the tensor constraint  $\Gamma = (B \otimes A)$ , where  $A \in \mathcal{O}_{p,\tilde{p}}$  and  $B \in \mathcal{O}_{q,\tilde{q}}$ . The following theorem established in Ye (2005) characterizes some useful properties of the solutions of the minimization problem (3). In the rest of discussion,  $P_M$  denotes the orthogonal projection matrix onto  $\text{span}(M)$  and  $Q_M = I - P_M$ .

**Theorem 2.1.** (Ye, 2005) Let  $\hat{A}, \hat{B}, \{\hat{U}_i\}_{i=1}^n$  constitutes a minimizer for (3) under the dimensionality  $(\tilde{p}, \tilde{q})$ . Then,

- (a)  $\hat{U}_i = \hat{A}^T(X_i - \bar{X})\hat{B}$ .
- (b)  $\{\hat{A}, \hat{B}\}$  is the maximizer of  $\frac{1}{n} \sum_{i=1}^n \|A^T(X_i - \bar{X})B\|_F^2$ .
- (c)  $\hat{A}$  consists of the leading  $\tilde{p}$  eigenvectors of  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})P_{\hat{B}}(X_i - \bar{X})^T$ , and  $\hat{B}$  consists of the leading  $\tilde{q}$  eigenvectors of  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^T P_{\hat{A}}(X_i - \bar{X})$ .

Similarly, we can define a population version of (3):  $E\|(X - \mu) - AUB^T\|_F^2$ , and the corresponding minimizer should follow Theorem 2.1 such that the minimizer over  $A \in \mathcal{O}_{p,\tilde{p}}$  and  $B \in \mathcal{O}_{q,\tilde{q}}$ , is equivalent to the maximizer of the maximization problem:

$$\arg\max_{A \in \mathcal{O}_{p,\tilde{p}}, B \in \mathcal{O}_{q,\tilde{q}}} E\|A^T(X - \mu)B\|_F^2 = \arg\max_{A \in \mathcal{O}_{p,\tilde{p}}, B \in \mathcal{O}_{q,\tilde{q}}} \text{trace}\{(B \otimes A)^T \Sigma (B \otimes A)\}, \quad (5)$$

where  $\Sigma = \text{Cov}(\text{vec}(X))$ . The following proposition gives the existence of the solution.

**Proposition 2.2.** For a fixed but arbitrary positive semi-definite matrix  $\Sigma$  of size  $pq \times pq$ , solution(s) to the maximization problem (5) exists.

Note that we do not need the model assumption (1) for Proposition 2.2. Also note that Proposition 2.2 applies to problem (3) as well by replacing  $\Sigma$  with its sample estimate  $S_n$ , the sample covariance matrix of  $\{\text{vec}(X_i)\}_{i=1}^n$ , and by rephrasing the maximization problem into the equivalent minimization problem. With the existence of the maximizer in (5) we can formally define the tensor principal components and the MPCA subspace.

**Definition 2.3.** For a pre-specified dimensionality  $(\tilde{p}, \tilde{q})$ , let  $\{A, B\}$  be the unique solution to the maximization problem (5), where  $A$  and  $B$  can be expressed in their columns as  $A = [a_1, \dots, a_{\tilde{p}}]$  and  $B = [b_1, \dots, b_{\tilde{q}}]$ . We call  $\{b_j \otimes a_i : 1 \leq i \leq \tilde{p}, 1 \leq j \leq \tilde{q}\}$  the tensor principal components, and  $\text{span}(B \otimes A)$  the MPCA subspace of dimensionality  $(\tilde{p}, \tilde{q})$ .

Using similar arguments as in Theorem 2.1 (c), we have that  $A$  and  $B$  consist of the leading  $\tilde{p}$  and  $\tilde{q}$  eigenvectors of  $E[(X - \mu)P_B(X - \mu)^T]$  and  $E[(X - \mu)^T P_A(X - \mu)]$ , respectively. Since

$$E[(X - \mu)P_B(X - \mu)^T] = \sum_{j=1}^{\tilde{q}} E[(X - \mu)(b_j b_j^T)(X - \mu)^T] = \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p), \quad (6)$$

$$E[(X - \mu)^T P_A(X - \mu)] = \sum_{i=1}^{\tilde{p}} E[(X - \mu)^T (a_i a_i^T)(X - \mu)] = \sum_{i=1}^{\tilde{p}} (a_i \otimes I_q)^T \Sigma (a_i \otimes I_q), \quad (7)$$

equivalently,  $\{A, B\}$  consist of the leading solutions of the system of stationary equations

$$\begin{aligned} \left( \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p) \right) a_i &= \lambda_i a_i, \quad i = 1, \dots, \tilde{p}, \\ \left( \sum_{i=1}^{\tilde{p}} (I_q \otimes a_i)^T \Sigma (I_q \otimes a_i) \right) b_j &= \xi_j b_j, \quad j = 1, \dots, \tilde{q}, \end{aligned}$$

over  $A \in \mathcal{O}_{p \times \tilde{p}}$  and  $B \in \mathcal{O}_{q \times \tilde{q}}$ , where the ordering is determined by the corresponding eigenvalues  $\lambda_1 \geq \dots \geq \lambda_{\tilde{p}} \geq 0$  and  $\xi_1 \geq \dots \geq \xi_{\tilde{q}} \geq 0$ .

**Remark 2.4.** Obviously  $\lambda_i$ 's,  $a_i$ 's,  $\xi_j$ 's and  $b_j$ 's depend on  $\Sigma$ . Besides such dependence, they also depend on the dimensionality  $(\tilde{p}, \tilde{q})$ . A more precise notation for them should be  $\lambda_i(\Sigma, \tilde{p}, \tilde{q})$ ,  $a_i(\Sigma, \tilde{p}, \tilde{q})$ ,  $\xi_j(\Sigma, \tilde{p}, \tilde{q})$  and  $b_j(\Sigma, \tilde{p}, \tilde{q})$ . However, for notation simplicity, we use  $\lambda_i$ ,  $a_i$ ,  $\xi_j$  and  $b_j$ , unless we want to emphasize on their dependence on  $(\Sigma, \tilde{p}, \tilde{q})$ .

From Remark 2.4, for any fixed  $(\tilde{p}, \tilde{q})$ , we could define the sample analogues  $\{\hat{A}, \hat{B}\}$ ,  $\hat{\lambda}_i$ 's, and  $\hat{\xi}_j$ 's by replacing  $\Sigma$  with the sample covariance matrix  $S_n$ . In the rest of the discussion, with pre-specified dimensionality  $(\tilde{p}, \tilde{q})$ , we denote the solution of (5) by  $\{A, B\}$  and the population tensor principal components by  $(B \otimes A)$ , and the corresponding sample analogues by  $\{\hat{A}, \hat{B}\}$  and  $(\hat{B} \otimes \hat{A})$ . Finding principal components in conventional PCA is equivalent to an eigenvalue-problem. However, there is no explicit solution of  $\{\hat{A}, \hat{B}\}$

for MPCA; therefore, an algorithm was proposed. The GLRAM algorithm of Ye (2005) to obtain  $\{\hat{A}, \hat{B}\}$  is summarized below.

**GLRAM (Ye, 2005):** Given a random initial  $A^{(0)} \in \mathcal{O}_{p \times \tilde{p}}$ . For  $k = 1, 2, \dots$ ,

1. Obtain the maximizer  $B^{(k+1)} = \arg \max_{B \in \mathcal{O}_{q \times \tilde{q}}} \frac{1}{n} \sum_{i=1}^n \|A^{(k)T}(X_i - \bar{X})B\|_F^2$ .
2. Obtain the maximizer  $A^{(k+1)} = \arg \max_{A \in \mathcal{O}_{p \times \tilde{p}}} \frac{1}{n} \sum_{i=1}^n \|A^T(X_i - \bar{X})B^{(k+1)}\|_F^2$ .
3. Repeat Steps 1-2 until there is no significant difference between  $\frac{1}{n} \sum_{i=1}^n \|A^{(k)T}(X_i - \bar{X})B^{(k)}\|_F^2$  and  $\frac{1}{n} \sum_{i=1}^n \|A^{(k+1)T}(X_i - \bar{X})B^{(k+1)}\|_F^2$ . Output  $\{\hat{A}, \hat{B}\} = \{A^{(k+1)}, B^{(k+1)}\}$ .

For any fixed  $A^{(k)}$  or  $B^{(k+1)}$ , the optimization problems in Steps 1 and 2 are the usual eigenvalue-problems of sizes  $p$  and  $q$ , respectively. Hence,  $A^{(k+1)}$  and  $B^{(k+1)}$  can be easily obtained. Moreover, the algorithm ensures the quantity  $\frac{1}{n} \sum_{i=1}^n \|A^{(k)T}(X_i - \bar{X})B^{(k)}\|_F^2$  to be monotonically increasing as  $k$  increases and, hence, the solution must exist since  $\frac{1}{n} \sum_{i=1}^n \|A^{(k)T}(X_i - \bar{X})B^{(k)}\|_F^2$  is bounded above by  $\frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}\|_F^2$ . Because GLRAM can only find a local maximum (depends on the chosen random initial  $A^{(0)}$ ), multiple random initials are suggested by Ye (2005) to ensure the global maximum. In contrast to this suggestion, we propose to use the leading  $\tilde{p}$  eigenvectors of  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$  as an initial of  $A^{(0)}$ .

We observe that hierarchical nesting structure may not exist for MPCA. Precisely, if  $(\tilde{p}', \tilde{q}') \leq (\tilde{p}, \tilde{q})$  with the corresponding solution pairs  $\{\hat{A}', \hat{B}'\}$  and  $\{\hat{A}, \hat{B}\}$ , respectively, there is no guarantee that  $\text{span}(\hat{A}') \subseteq \text{span}(\hat{A})$ , nor  $\text{span}(\hat{B}') \subseteq \text{span}(\hat{B})$ . In the population level, however, there certainly exist relationships between the target subspaces and the MPCA subspaces prescribed by the optimization problem (5).

**Proposition 2.5.** *Assume model (1) and let  $\{A, B\}$  be the solution pair to the maximization problem (5) under dimensionality  $(\tilde{p}, \tilde{q})$ .*

- (a) *If  $\tilde{p} \geq p_0$  and  $\tilde{q} \geq q_0$ , then  $\text{span}(A) \supseteq \text{span}(A_0)$  and  $\text{span}(B) \supseteq \text{span}(B_0)$ .*
- (b) *If  $\tilde{p} < p_0$  and  $\tilde{q} \geq q_0$ , then  $\text{span}(A) \subsetneq \text{span}(A_0)$  and  $\text{span}(B) \supseteq \text{span}(B_0)$ .*
- (c) *If  $\tilde{p} \geq p_0$  and  $\tilde{q} < q_0$ , then  $\text{span}(A) \supseteq \text{span}(A_0)$  and  $\text{span}(B) \subsetneq \text{span}(B_0)$ .*



(d) If  $\tilde{p} < p_0$  and  $\tilde{q} < q_0$ , then  $\text{span}(A) \subsetneq \text{span}(A_0)$  and  $\text{span}(B) \subsetneq \text{span}(B_0)$ .

Even though there is no general hierarchical nesting structure for MPCA subspaces, Proposition 2.5 ensures the existence of a specific nesting structure, which the extracted MPCA subspace is a proper subspace of the target subspace if the dimension is under-specified, and contains the target subspace if the dimension is over-specified. It also implies that MPCA indeed searches the true target subspace  $\text{span}(B_0 \otimes A_0)$  when  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$  is correctly specified. As a result, these arguments provide a justification of using  $\text{span}(\hat{B} \otimes \hat{A})$  in the sample level for subsequent statistical analysis.

## 2.2 Connection with (2D)<sup>2</sup>PCA and conventional PCA

The (2D)<sup>2</sup>PCA is another method to extract basis for tensor objects. For a given dimensionality  $(\tilde{p}, \tilde{q})$ , the population (2D)<sup>2</sup>PCA components  $A^* = [a_1^*, \dots, a_{\tilde{p}}^*]$  and  $B^* = [b_1^*, \dots, b_{\tilde{q}}^*]$  are defined to be the leading  $\tilde{p}$  and  $\tilde{q}$  eigenvectors of  $E[(X - \mu)(X - \mu)^T]$  and  $E[(X - \mu)^T(X - \mu)]$  with the corresponding eigenvalues  $\{\lambda_i^* : 1 \leq i \leq \tilde{p}\}$  and  $\{\xi_j^* : 1 \leq j \leq \tilde{q}\}$ . The sample analogues, denoted by  $\hat{A}^*$ ,  $\hat{B}^*$ ,  $\hat{\lambda}_i^*$ , and  $\hat{\xi}_j^*$  are similarly defined to be the leading eigenvectors and eigenvalues of  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$  and  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^T(X_i - \bar{X})$ . The following proposition states a connection between the (2D)<sup>2</sup>PCA and MPCA in the population level.

**Proposition 2.6.** *Assume model (1) and that  $\{\lambda_i^* : 1 \leq i \leq p_0\}$  and  $\{\xi_j^* : 1 \leq j \leq q_0\}$  are simple roots.*

- (a) *If  $\tilde{q} \geq q_0$ , then MPCA and (2D)<sup>2</sup>PCA share the same leading  $(p_0 \wedge \tilde{p})$  eigenvectors, i.e.,  $a_i = a_i^*$  for  $i = 1, \dots, (p_0 \wedge \tilde{p})$ . Moreover,  $\lambda_i^* - \lambda_i = (q - \tilde{q})\sigma^2$  for  $i = 1, \dots, \tilde{p}$ .*
- (b) *If  $\tilde{p} \geq p_0$ , then MPCA and (2D)<sup>2</sup>PCA share the same leading  $(q_0 \wedge \tilde{q})$  eigenvectors, i.e.,  $b_j = b_j^*$  for  $j = 1, \dots, (q_0 \wedge \tilde{q})$ . Moreover,  $\xi_j^* - \xi_j = (p - \tilde{p})\sigma^2$  for  $j = 1, \dots, \tilde{q}$ .*

When the dimension  $(\tilde{p}, \tilde{q})$  is adequate, Proposition 2.6 implies that (2D)<sup>2</sup>PCA and MPCA, in the population level, actually target the same subspace  $\text{span}(B_0 \otimes A_0)$  under model (1). However, there is no guarantee that the extracted bases  $\{\hat{A}^*, \hat{B}^*\}$  of (2D)<sup>2</sup>PCA also maximize the sample version of (5). Though, under the setting of Proposition 2.6,

$E[(X - \mu)P_B(X - \mu)^T]$  and  $E[(X - \mu)(X - \mu)^T]$  have the same leading eigenvectors, we expect an efficiency gain in using  $E[(X - \mu)P_B(X - \mu)^T]$ , since it is less noise-contaminated than  $E[(X - \mu)(X - \mu)^T]$ . A rigorous proof of efficiency gain is provided in Section 3.

**Remark 2.7.** *From Proposition 2.6, it is suggested to select  $(\tilde{p}, \tilde{q})$  through (2D)<sup>2</sup>PCA, since the dimension of  $A_0$  and  $B_0$  are not known before being estimated. A formal statistical test is provided in Section 2.3.*

There is also a connection between MPCA and conventional PCA. Under model (1), without considering the random noise  $\varepsilon$ ,  $\text{vec}(X - \mu)$  belongs to  $\text{span}(B_0 \otimes A_0)$ , which is the target subspace of MPCA. Observe also that

$$\begin{aligned}\Sigma &= (B_0 \otimes A_0)T(B_0 \otimes A_0)^T + \sigma^2 I_m \\ &= (B_0 \otimes A_0)(T + \sigma^2 I_{m_0})(B_0 \otimes A_0)^T + \sigma^2 Q_{B_0 \otimes A_0},\end{aligned}\tag{8}$$

where  $Q_{B_0 \otimes A_0} = Q_{B_0} \otimes P_{A_0} + P_{B_0} \otimes Q_{A_0} + Q_{B_0} \otimes Q_{A_0}$  is the projection matrix onto the complement of  $\text{span}(B_0 \otimes A_0)$ . It should be noted that the matrix  $T$  is not necessarily a diagonal matrix. Hence,  $(B_0 \otimes A_0)$  is not the same with the conventional PCA components in general. If we further diagonalize  $T = G D G^T$  with  $D$  being a diagonal matrix of size  $m_0 \times m_0$ , we have the following factorization:

$$\Sigma = [\Gamma, \Gamma_\perp] \begin{bmatrix} D + \sigma^2 I_{m_0} & 0 \\ 0 & \sigma^2 I_{m-m_0} \end{bmatrix} [\Gamma, \Gamma_\perp]^T,$$

where  $\Gamma = (B_0 \otimes A_0)G$  and  $\Gamma_\perp$  is an orthonormal basis for the orthogonal complement of  $\text{span}(\Gamma)$ . Consequently, the conventional PCA uses  $\Gamma = (B_0 \otimes A_0)G$  as coordinate system for a compressed representation for  $\text{vec}(X - \mu)$ , while the MPCA uses  $(B_0 \otimes A_0)$ . Notice that  $\text{span}(B_0 \otimes A_0) = \text{span}(\Gamma)$  provided that  $T$  in (8) is of full rank. In summary, MPCA and conventional PCA use the same subspace for compressed data representation. However, MPCA requires less parameters (see the following remark) to specify the low-dimensional subspace than the conventional approach.

**Remark 2.8.** *The number of free parameters required for MPCA is  $p_0 p - \frac{1}{2} p_0 (p_0 + 1) + q_0 q - \frac{1}{2} q_0 (q_0 + 1)$ , which is relatively small in contrast with the number of free parameters*

required for conventional PCA:  $p_0 q_0 p q - \frac{1}{2} p_0 q_0 (p_0 q_0 + 1)$ . It is the adoption of  $(B_0 \otimes A_0)$  for the sake of parsimony, which is one of the purposes of using MPCA. The following table gives the numbers of parameters needed to specify an orthonormal basis for a subspace of dimensionality  $p_0 \times q_0$  within a space of dimensionality  $p \times q = 100$ . We fix  $(p, q, p_0) = (10, 10, 5)$  and let  $q_0$  vary.

| $q_0$ | 1   | 2   | 3    | 4    | 5    |
|-------|-----|-----|------|------|------|
| MPCA  | 44  | 52  | 59   | 65   | 70   |
| PCA   | 485 | 945 | 1380 | 1790 | 2175 |

Table 1: Numbers of required free parameters at  $(p, q, p_0) = (10, 10, 5)$ .

We remind the reader that there is no obvious ordering relationship between the MPCA components and conventional PCA components. This can be seen in a simple example when  $T = \text{Cov}(\text{vec}(U)) = \text{diag}(\text{vec}(C))$ , where  $C$  is a matrix with  $C_{ij} = \text{Var}(U_{ij})$ . For the case of uncorrelated  $U_{ij}$ 's,  $T$  is diagonal, and hence, the conventional PCA and the MPCA share the same eigenvectors. The leading  $p_0 q_0$  eigenvalues of the conventional PCA are  $\{C_{ij} + \sigma^2 : 1 \leq i \leq p_0, 1 \leq j \leq q_0\}$ , which have a natural ordering depending on the values of  $C_{ij}$ 's. On the other hand, the leading eigenvalues of MPCA at  $(p_0, q_0)$  are derived to be  $C_{i\cdot} = \sum_{j=1}^{q_0} C_{ij}$ ,  $i = 1, \dots, p_0$ , and  $C_{\cdot j} = \sum_{i=1}^{p_0} C_{ij}$ ,  $j = 1, \dots, q_0$ , where the ordering depends on the column sums and row sums of  $C_{ij}$ 's. Therefore, even if we pick  $a_i$  and  $b_j$  from leading eigenvectors of  $A$  and  $B$ , there is no guarantee that, when paired together,  $b_j \otimes a_i$  is on the top list of leading eigenvectors of the conventional PCA.

## 2.3 Selection of dimensionality

This section is devoted to the selection of the dimensionality  $(\tilde{p}, \tilde{q})$ . Similar to the conventional PCA, we propose that the dimension is determined by the explained variance, as a popular method in conventional PCA. First we define the cumulative variance, which is a measure of the total variance of the tensor objects projected onto MPCA subspace.

**Definition 2.9.** Let  $\{A, B\}$  be a solution pair to the problem (5). We call the quantity  $\Phi(\tilde{p}, \tilde{q}) = E\|A^T(X - \mu)B\|_F^2$  the cumulative variance for  $X$  at rank- $(\tilde{p}, \tilde{q})$ , and the quantity

$$\rho(\tilde{p}, \tilde{q}) = \frac{\Phi(\tilde{p}, \tilde{q})}{\Phi(p, q)}$$

the explained percentage of total variance of  $X$  at rank- $(\tilde{p}, \tilde{q})$ . Note that  $\Phi(p, q) = E\|X - \mu\|_F^2$ . The corresponding sample analogues are defined to be  $\hat{\Phi}(\tilde{p}, \tilde{q}) = \frac{1}{n} \sum_{i=1}^n \|\hat{A}^T(X_i - \bar{X})\hat{B}\|_F^2$ ,  $\hat{\Phi}(p, q) = \frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}\|_F^2$  and

$$\hat{\rho}(\tilde{p}, \tilde{q}) = \frac{\hat{\Phi}(\tilde{p}, \tilde{q})}{\hat{\Phi}(p, q)}.$$

**Remark 2.10.** Note that  $\Phi(\tilde{p}, \tilde{q}) > \Phi(\tilde{p}', \tilde{q}')$  does not necessarily imply  $\text{span}(A_{\tilde{p}} \otimes B_{\tilde{q}}) \supset \text{span}(A_{\tilde{p}'} \otimes B_{\tilde{q}'})$ . Similar phenomenon can be observed on the cumulative distribution function. For instance, in a 2-dimensional c.d.f  $F$ , the phenomenon “ $F(x_1, x_2) > F(x'_1, x'_2)$ ” does not imply  $\{(u, v) : u \leq x_1, v \leq x_2\} \supset \{(u, v) : u \leq x'_1, v \leq x'_2\}$ .

From the description below Definition 2.3, we have  $\Phi(\tilde{p}, \tilde{q}) = \sum_{i=1}^{\tilde{p}} \lambda_i = \sum_{j=1}^{\tilde{q}} \xi_j$  and  $\hat{\Phi}(\tilde{p}, \tilde{q}) = \sum_{i=1}^{\tilde{p}} \hat{\lambda}_i = \sum_{j=1}^{\tilde{q}} \hat{\xi}_j$ . Note that  $\lambda_i$  and  $\xi_j$ , as well as  $\hat{\lambda}_i$  and  $\hat{\xi}_j$ , depend on the specified dimensionality  $(\tilde{p}, \tilde{q})$ . Also note that  $\Phi(\tilde{p}, \tilde{q}) \leq \Phi(p, q)$  always holds. Thus,  $\rho(\tilde{p}, \tilde{q}) \leq 1$  and is used as a measure of adequacy for MPCA at dimensionality  $(\tilde{p}, \tilde{q})$ . Specifically, for a given  $\rho_0 \in (0, 1)$ , consider the hypothesis test:

$$H_0 : \rho(\tilde{p}, \tilde{q}) \leq \rho_0 \quad v.s. \quad H_1 : \rho(\tilde{p}, \tilde{q}) > \rho_0. \quad (9)$$

A rejection of  $H_0$  then indicates the chosen dimensionality  $(\tilde{p}, \tilde{q})$  satisfies the condition that  $\rho(\tilde{p}, \tilde{q})$  reaches the required level of explained variance at a certain confidence. To perform the test, a reference distribution for the sample analogue  $\hat{\rho}(\tilde{p}, \tilde{q})$  is required. We derive the asymptotic distribution of  $\sqrt{n}(\hat{\rho}(\tilde{p}, \tilde{q}) - \rho(\tilde{p}, \tilde{q}))$  in Section 3, which can be used to construct the rejection region of the test.

### 3 Asymptotic properties for MPCA

In this section, we investigate the asymptotic behavior of MPCA. Without loss of generality, we assume  $\mu = 0$  to simplify the notations in the rest of discussion. It then implies

$\Sigma = E[\text{vec}(X)\text{vec}(X)^T]$  and the population kernel matrices of MPCA at dimensionality  $(\tilde{p}, \tilde{q})$  can be simplified to be  $E[X P_B X^T]$  and  $E[X^T P_A X]$ . Note also that the population kernel matrices of (2D)<sup>2</sup>PCA reduce to  $E[XX^T]$  and  $E[X^T X]$  in this situation.

Let  $S_n$  be the sample covariance matrix of  $\{\text{vec}(X_i)\}_{i=1}^n$ , where  $X_i$ 's are iid observations with finite second moments following model (1). By the central limit theorem, we have

$$\sqrt{n}(S_n - \Sigma) \xrightarrow{d} N, \quad (10)$$

where  $\text{vec}(N)$  is an  $m^2$ -variate normal with zero-mean and covariance matrix  $\Sigma_N = \text{Cov}(\text{vec}(X) \otimes \text{vec}(X))$ . If  $\text{vec}(X)$  is further assumed to be normally distributed, then  $S_n$  follows a Wishart distribution and  $\Sigma_N$  is derived to be (Anderson, 1963)

$$\Sigma_N = (I_{m^2} + K_{m,m})(\Sigma \otimes \Sigma), \quad (11)$$

where  $K_{\ell,k} = \sum_{i=1}^{\ell} \sum_{j=1}^k H_{ij} \otimes H_{ij}^T$  is the commutation matrix, and  $H_{ij}$  is an  $\ell \times k$  matrix with one in the  $(i, j)^{\text{th}}$  entry and zeros elsewhere. Some important properties involving commutation matrix are listed here (Magnus and Neudecker, 1979). Let  $M_1 \in \Re^{a_1 \times b_1}$  and  $M_2 \in \Re^{a_2 \times b_2}$  be two arbitrary matrices. Then,  $K_{a_1, b_1} = K_{b_1, a_1}^T$ ,  $K_{a_1, b_1} K_{b_1, a_1} = I_{a_1 b_1}$ ,  $K_{a_1, b_1} = I_{a_1}$  if  $b_1 = 1$ ,  $\text{vec}(M_1^T) = K_{a_1, b_1} \text{vec}(M_1)$ , and  $(M_2 \otimes M_1) = K_{a_2, a_1} (M_1 \otimes M_2) K_{b_1, b_2}$ . These properties will be repeatedly used in the discussion of asymptotic theory without further reference. We note that, unless explicitly specified, the asymptotic properties derived in this section does not rely on the normality of  $\text{vec}(X)$ .

### 3.1 Asymptotic distributions for principal components, projections, cumulative variance and explained variance in MPCA

We first state the weak convergence of the cumulative variances and the tensor principal components of MPCA. The limiting distributions for projections and explained variance are direct applications of delta method.

**Theorem 3.1.** *Assume model (1) and, for any fixed  $(\tilde{p}, \tilde{q})$  with  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ , the leading  $\tilde{p}$  eigenvalues  $\lambda_i(\Sigma, \tilde{p}, \tilde{q})$ 's and the leading  $\tilde{q}$  eigenvalues  $\xi_j(\Sigma, \tilde{p}, \tilde{q})$ 's of MPCA are simple roots.*

(a) For  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ ,<sup>1</sup> we have the limiting distribution

$$\sqrt{n} \left( \begin{bmatrix} \hat{\Phi}(\tilde{p}, \tilde{q}) \\ \hat{\Phi}(p, q) \end{bmatrix} - \begin{bmatrix} \Phi(\tilde{p}, \tilde{q}) \\ \Phi(p, q) \end{bmatrix} \right) \xrightarrow{d} \begin{bmatrix} D_{\Phi(\tilde{p}, \tilde{q})} \\ \text{vec}(I_m)^T \end{bmatrix} \text{vec}(N), \quad (12)$$

where  $D_{\Phi(\tilde{p}, \tilde{q})} = \frac{\partial \Phi(\tilde{p}, \tilde{q})}{\partial \text{vec}(\Sigma)}$  and its explicit expression is given in Lemma 3.2.

(b) For  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ ,<sup>2</sup> we have the limiting distribution

$$\sqrt{n} \left( \begin{bmatrix} \text{vec}(\hat{A}) \\ \text{vec}(\hat{B}) \end{bmatrix} - \begin{bmatrix} \text{vec}(A) \\ \text{vec}(B) \end{bmatrix} \right) \xrightarrow{d} D_{H_{\tilde{p}, \tilde{q}}} \text{vec}(N), \quad (13)$$

where

$$D_{H_{\tilde{p}, \tilde{q}}} = \left[ \left( \frac{\partial a_1}{\partial \text{vec}(\Sigma)} \right)^T, \dots, \left( \frac{\partial a_{\tilde{p}}}{\partial \text{vec}(\Sigma)} \right)^T, \left( \frac{\partial b_1}{\partial \text{vec}(\Sigma)} \right)^T, \dots, \left( \frac{\partial b_{\tilde{q}}}{\partial \text{vec}(\Sigma)} \right)^T \right]^T.$$

When  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ ,  $D_{H_{p_0, q_0}}$  has an explicit expression, which is given in Lemma 3.2.

**Lemma 3.2.** Assume the model (1).

(a) For  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ , we have

$$D_{\Phi(\tilde{p}, \tilde{q})} = \text{vec}(P_{B \otimes A})^T. \quad (14)$$

(b) When  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ , for  $i = 1, \dots, p_0$  and  $j = 1, \dots, q_0$ , we have

$$\frac{\partial a_i}{\partial \text{vec}(\Sigma)} = \left\{ a_i \otimes \text{vec}(P_{B_0}) \otimes (\lambda_i I_p - E[X P_{B_0} X^T])^+ \right\}^T (K_{p,q} \otimes I_{pq}) \quad (15)$$

$$\frac{\partial b_j}{\partial \text{vec}(\Sigma)} = \left\{ b_j \otimes \text{vec}(P_{A_0}) \otimes (\xi_j I_q - E[X^T P_{A_0} X])^+ \right\}^T (I_{pq} \otimes K_{p,q}), \quad (16)$$

where, for a given matrix  $M$ ,  $M^+$  denotes its Moore-Penrose generalized inverse.

It can be seen from Lemma 3.2 that, when  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ , the asymptotic distribution of  $\hat{A}$  depends on  $B$  only through  $\text{span}(B) = \text{span}(B_0)$ , and the asymptotic distribution of  $\hat{B}$  depends on  $A$  only through  $\text{span}(A) = \text{span}(A_0)$ . We are now on the position to obtain the asymptotic normality of the projection matrix onto MPCA subspace  $P_{\hat{B} \otimes \hat{A}}$  and the explained variance  $\hat{\rho}(\tilde{p}, \tilde{q})$  in the following corollaries.

---

<sup>1</sup>For  $\tilde{p} > p_0$  (or  $\tilde{q} > q_0$ , resp.)  $E[XX^T]$  (or  $E[X^T X]$ , resp.) has multiple roots from the  $(p_0 + 1)^{\text{th}}$  (or  $(q_0 + 1)^{\text{th}}$ , resp.) eigenvalue and beyond.

<sup>2</sup>For either  $\tilde{p} > p_0$ , or  $\tilde{q} > q_0$ , the  $(p_0 + 1)^{\text{th}}$ , or  $(q_0 + 1)^{\text{th}}$ , tensor principal components are not uniquely determined due to multiple characteristic roots.

**Corollary 3.3.** *Under the same assumptions of Theorem 3.1. For  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ , we have the limiting distribution of the projection matrix onto MPCA subspace*

$$\sqrt{n} \text{vec}(P_{\hat{B} \otimes \hat{A}} - P_{B \otimes A}) \xrightarrow{d} D_{P_{B \otimes A}} \text{vec}(N), \quad (17)$$

where  $D_{P_{B \otimes A}} = \frac{\partial \text{vec}(P_{B \otimes A})}{\partial \text{vec}(\Sigma)}$ . When  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ ,  $D_{P_{B_0 \otimes A_0}}$  has the explicit expression

$$\begin{aligned} & (I_{m^2} + K_{m,m}) \times \\ & \left\{ \sum_{i=1}^{p_0} (K_{q,p} \otimes I_{pq}) (P_{a_i} \otimes [\text{vec}(P_{B_0}) \text{vec}(P_{B_0})^T] \otimes \{\lambda_i I_p - E[X P_{B_0} X^T]\}^+) (K_{p,q} \otimes I_{pq}) \right. \\ & \left. + \sum_{j=1}^{q_0} (I_{pq} \otimes K_{q,p}) (P_{b_j} \otimes [\text{vec}(P_{A_0}) \text{vec}(P_{A_0})^T] \otimes \{\xi_j I_q - E[X^T P_{A_0} X]\}^+) (I_{pq} \otimes K_{p,q}) \right\}. \end{aligned}$$

**Corollary 3.4.** *Under the same assumptions of Theorem 3.1. For  $\tilde{p} \leq p_0$  and  $\tilde{q} \leq q_0$ , we have the limiting distribution of the explained variance*

$$\sqrt{n}(\hat{\rho}(\tilde{p}, \tilde{q}) - \rho(\tilde{p}, \tilde{q})) \xrightarrow{d} N(0, \sigma_{\rho(\tilde{p}, \tilde{q})}^2), \quad (18)$$

where  $\sigma_{\rho(\tilde{p}, \tilde{q})}^2$  is defined to be

$$\left( \frac{1}{\Phi(p, q)} D_{\Phi(\tilde{p}, \tilde{q})} - \frac{\Phi(\tilde{p}, \tilde{q})}{\Phi(p, q)^2} \text{vec}(I_m)^T \right) \Sigma_N \left( \frac{1}{\Phi(p, q)} D_{\Phi(\tilde{p}, \tilde{q})} - \frac{\Phi(\tilde{p}, \tilde{q})}{\Phi(p, q)^2} \text{vec}(I_m)^T \right)^T. \quad (19)$$

Corollary 3.4 is the cornerstone of our asymptotic test for hypothesis (9). Before practical implementation of the test, however, we need a consistent estimator of  $\sigma_{\rho(\tilde{p}, \tilde{q})}^2$ . Note that the asymptotic covariance  $\Sigma_N$  can be empirically estimated by

$$\hat{\Sigma}_{N,1} = \frac{1}{n} \sum_{i=1}^n \text{vec}(\text{vec}(X_i - \bar{X}) \text{vec}(X_i - \bar{X})^T - S_n) \text{vec}(\text{vec}(X_i - \bar{X}) \text{vec}(X_i - \bar{X})^T - S_n)^T.$$

Moreover, if  $\text{vec}(X)$  is normally distributed, we can also estimate  $\Sigma_N$  by

$$\hat{\Sigma}_{N,2} = (I_{m^2} + K_{m,m})(S_n \otimes S_n)$$

based on (11). Consequently, the asymptotic variance  $\sigma_{\rho(\tilde{p}, \tilde{q})}^2$  is estimated by

$$\hat{\sigma}_{\rho(\tilde{p}, \tilde{q})}^2 = \left( \frac{1}{\hat{\Phi}(p, q)} \hat{D}_{\Phi(\tilde{p}, \tilde{q})} - \frac{\hat{\Phi}(\tilde{p}, \tilde{q})}{\hat{\Phi}(p, q)^2} \text{vec}(I_m)^T \right) \hat{\Sigma}_{N,i} \left( \frac{1}{\hat{\Phi}(p, q)} \hat{D}_{\Phi(\tilde{p}, \tilde{q})} - \frac{\hat{\Phi}(\tilde{p}, \tilde{q})}{\hat{\Phi}(p, q)^2} \text{vec}(I_m)^T \right)^T$$

for  $i = 1, 2$  (depends on the normality of  $\text{vec}(X)$  or not), where  $\hat{D}_{\Phi(\tilde{p}, \tilde{q})} = \text{vec}(P_{\hat{B} \otimes \hat{A}})^T$ .

The consistency of  $\hat{\sigma}_{\rho(\tilde{p}, \tilde{q})}^2$  is a direct consequence by standard arguments. These facts enable us to construct an approximate level  $\alpha$  test to determine the dimensionality  $(\tilde{p}, \tilde{q})$ .

**Theorem 3.5.** *Assume the conditions of Theorem 3.1 and  $(\tilde{p}, \tilde{q}) \leq (p_0, q_0)$ . For the hypothesis (9), an approximated level  $\alpha$  test is to reject  $H_0$  if*

$$\hat{\rho}(\tilde{p}, \tilde{q}) > \rho_0 + \frac{\hat{\sigma}_{\rho(\tilde{p}, \tilde{q})}}{\sqrt{n}} z_\alpha, \quad (20)$$

where  $z_\alpha$  is the upper  $\alpha$  quantile of the standard normal.

### 3.2 Asymptotic efficiency

MPCA and  $(2D)^2$ PCA actually target the same basis when  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ . Intuitively, we are in favor of MPCA since its kernel matrices are less noise-contaminated than the ones of  $(2D)^2$ PCA as mentioned previously. The following theorem proves that MPCA is indeed asymptotically more efficient than  $(2D)^2$ PCA, wherein aCov denotes the asymptotic covariance.

**Theorem 3.6.** *Assume the conditions of Theorem 3.1 and the normality of  $\text{vec}(X)$ . Let  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$  and let  $(\hat{A}^*, \hat{B}^*)$  be the  $(2D)^2$ PCA components under  $(p_0, q_0)$ . Then,*

$$\text{aCov}(\text{vec}(P_{\hat{B}^* \otimes \hat{A}^*})) - \text{aCov}(\text{vec}(P_{\hat{B} \otimes \hat{A}})) \geq 0, \quad (21)$$

where the equality holds if and only if  $(p_0, q_0) = (p, q)$ .

Theorem 3.6 states that under model (1), MPCA is at most as disperse as  $(2D)^2$ PCA in estimating the dimension reduction subspace  $\text{span}(B_0 \otimes A_0)$ . The only case that we will gain nothing from MPCA over the  $(2D)^2$ PCA is when  $(p_0, q_0) = (p, q)$ . Note that the condition  $(p_0, q_0) = (p, q)$  implies that there is no room of dimension reduction at all and is probably of no interest in real applications. Consequently, Theorem 3.6 provides a justification of using MPCA.

## 4 Experimental study: the Olivetti Faces data set

We test and compare the performance of MPCA and conventional PCA on Olivetti Faces data set, which is available at <http://www.cs.nyu.edu/~roweis/data.html>. This data set consists of 400 gray scale (8 bits) face images of  $64 \times 64$  pixels. There exist different



facial expressions and/or views for each individual in this data set. A simulation experiment is designed as follows. 400 face images are randomly partitioned into a training set with size 100 and a test set with size 300. This 100-300 partition, where the training set is smaller than test set, is to reflect a scenario of using a small portion of data to train a basis set for the representation of the rest data in data archive.

Both MPCA and conventional PCA are applied on the 100 training images to produce image basis which is used to reconstruct the rest 300 test images. The average of the 100 training images, named mean face, has been subtracted from all the 400 images for PCA training and for test image reconstructions as well. The mean face is finally added to the reconstructions at the last stage to show the resulting images. 500 replicates of training-test partitions are performed to compare the mean test error, which is defined as the average of the Frobenius norm between the original images and the reconstructed images on test data set. The result is in Table 2. The mean test error for conventional PCA is more than seven times of that for MPCA; and the standard deviation is more than 12 times.

| Frobenius-Norm         | MPCA   | Conventional PCA |
|------------------------|--------|------------------|
| Mean ( $\times 10^5$ ) | 1.1346 | 8.6455           |
| SD ( $\times 10^2$ )   | 9.6398 | 120.39           |

Table 2: Test error comparison for MPCA and conventional PCA on the test images of Olivetti Faces data set. The error is defined as the Frobenius norm of two image matrices: original test image and its reconstruction using  $28 \times 28$  principal components.

In Figures 1-3, 40 test images are randomly chosen from the test set to show the visual performance of image reconstructions by these two PCA schemes. In MPCA, 28 row eigenvectors and 28 column eigenvectors, both with size 64, are used to generate 784 basis images, of which the 100 leading ones are shown in Figure 4. We remind the reader that the selection  $(\tilde{p}, \tilde{q}) = (28, 28)$  produces an  $\hat{\rho}(\tilde{p}, \tilde{q})$  value 0.968. Based on Theorem 3.5, a one-sided 95% confidence interval for  $\hat{\rho}(28, 28)$  is given by  $[0.967, 1]$ . We also show the variability pattern plots (Tu and Huang, 2011) in Figure 7. These plots present the average

variations (absolute values) of the eigenvectors for the bootstrap re-sampled data, from those eigenvectors for the original data. The horizontal and vertical indices refer to the eigenvector indices for re-sampled and original data. The indices of eigenvectors are sorted by eigenvalues. The variations are presented by colors from dark blue for perfect matched, to dark red for extremely deviated. Usually, eigenvectors with distinct eigenvalues show deep blue on the diagonal and deep red on the off-diagonal. Eigenvectors with the same multiple root eigenvalue tend to be visualized by a cubic pattern on their correspondence indices. It can be seen that our choice of  $(\tilde{p}, \tilde{q}) = (28, 28)$  does not produce multiple roots, since the bootstrapped variability of the solutions at this selection is quite small.

In conventional PCA, 784 ( $= 28 \times 28$ ) eigenvectors (basis images) with size 4096 are used, of which the 100 leading ones are shown in Figure 5. Because of using 100 training images with average subtraction, there are at most 99 meaningful eigenvectors in the conventional PCA. The rest are randomly orthogonal eigenvectors with zero eigenvalue from the remaining subspace. In Figure 5, from top to bottom, we can see the images with clear facial shape to vague ones and a random image on the 100<sup>th</sup> one. On the other hand, MPCA tends to distribute the image characteristics to more basis elements which may allow for more local modification on the images.

In Figure 6, one particular image among the 40 test images is chosen to demonstrate the performance of these two methods. The top row shows the image reconstruction process for MPCA when more basis elements are added in, and the bottom shows for conventional PCA. The mean face is put in the first column and the target image in the 7<sup>th</sup> column as references. The right-most column shows the absolute values of projection scores on the leading 784 basis elements. It is clear that the conventional PCA concentrates on no more than 99 basis elements while the MPCA spreads out to much more basis elements. For MPCA, the image turns its view when 10 basis elements are used; the pupil turns to left when  $16 \times 16$  basis elements are used; the double eyelid and nostrils show up when  $22 \times 22$  basis elements are used; the facial curves become clear when  $28 \times 28$  basis elements are used. While we can observe the reconstruction progress by adding more basis elements for MPCA, we do not see much difference after 100 basis elements for conventional PCA. It is clear that MPCA performs better than conventional

PCA in reconstructing the test images from Table 2 and these figures.

## 5 Concluding discussions

PCA is a popular tool to reduce the dimensions for high dimensional data analysis; MPCA could be likely to serve the similar function for higher order tensor data sets. From this work, the statistical properties of MPCA become clear through the theoretical framework and the performance of MPCA is predictable through the asymptotic results. Most importantly, based on these asymptotic results, various hypothesis tests become feasible for subsequent analysis, including pattern recognition or classification. Our work, though technically theoretical, may construct a platform to expand the application potentials of MPCA.

The advantages of MPCA over conventional PCA on tensor structure data are evident in the Olivetti Faces data example. Therein, conventional PCA suffers seriously from the large  $m$  and small  $n$  problem such that there can be at most  $n - 1$  meaningful eigenvectors. This makes it unavoidable that all the data noises are still carried by the chosen principal components. Furthermore, too concentrated information in one component, which may not be good for pattern recognition or classification prediction. On the other hand, MPCA distributes the information to more components which may allow local modification in the process of image reconstruction, with even fewer free parameters. The key point for the good performance of MPCA is the data tensor structure. For practical purposes, the robustness of MPCA over model variety should be further investigated.

## References

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34, 122-148.
- De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21, 1253-1278.

- De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000b). On the best rank-1 and rank- $(R_1, R_2, \dots, R_N)$  approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21, 1324-1342.
- Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. *Statistics*, 18, 401-414.
- Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian J. Statistics*, 7, 65-81.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer, New York.
- Kolda, T.G. and Bader, B.W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3), 455-500.
- Li, B., Kim, M.K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Annals of Statistics*, 38, 1094-1121.
- Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19, 18-39.
- Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Annals of Statistics*, 7, 381-394.
- Sibson, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J. Roy. Statist. Soc.*, 41, 217-229.
- Tu, I. P. and Huang, H. C. (2011). An estimation on a covariance matrix when multiple roots exist. *manuscript*.
- Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics*, 9, 725-736.
- Yang, J., Zhang, D., Frangi, A.F. and Yang, J.Y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 131-137.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61, 167-191.

Zhang, D. and Zhou, Z. H. (2005). (2D)<sup>2</sup>PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 69, 224-231.

## Appendix

**Proof of Proposition 2.2.** In the maximization problem (5), the objective function is continuous and the feasible region  $\mathcal{O}_{q,\tilde{q}} \otimes \mathcal{O}_{p,\tilde{p}}$  is compact. (Both continuity and compactness are with respect to the topology induced by Frobenius norm.) Thus, solution(s) exists.  $\square$

**Proof of Proposition 2.5.** (a) Let  $[B, B_\perp]$  be a  $q \times q$  orthonormal matrix. Since  $B_0 \in \text{span}([B, B_\perp])$ , there exists  $\eta_1 \in \mathbb{R}^{\tilde{q} \times q_0}$  and  $\eta_2 \in \mathbb{R}^{(q-\tilde{q}) \times q_0}$  such that  $B_0 = B\eta_1 + B_\perp\eta_2$ . As  $B_0^T B_0 = I_{q_0}$ , we have  $\eta_1^T \eta_1 + \eta_2^T \eta_2 = I_{q_0}$ . Observe that

$$\begin{aligned}
& E \|A^T(X - \mu)B\|_F^2 \\
&= E \{ \text{trace}(A^T(A_0 U B_0^T) B B^T (B_0 U^T A_0^T) A) \} + E \{ \text{trace}(A^T \varepsilon B B^T \varepsilon^T A) \} \\
&= E \{ \text{trace}(A^T A_0 U \eta_1^T \eta_1 U^T A_0^T A) \} + \tilde{p} \tilde{q} \sigma^2 \\
&= E \{ \text{trace}(A^T A_0 U U^T A_0^T A) \} - E \{ \text{trace}(A^T A_0 U \eta_2^T \eta_2 U^T A_0^T A) \} + \tilde{p} \tilde{q} \sigma^2 \\
&\leq E \{ \text{trace}(A^T A_0 U U^T A_0^T A) \} + \tilde{p} \tilde{q} \sigma^2, \tag{A.1}
\end{aligned}$$

where the equality in (A.1) holds if and only if  $\eta_2 = 0$ , if and only if  $\eta_1^T \eta_1 = I_{q_0}$ . Thus, if  $\tilde{q} \geq q_0$ , such an  $\eta_1$  (with rank  $q_0$ ) exists to ensure the equality in (A.1). This implies  $B_0 = B\eta_1$  and, hence,  $B_0 \in \text{span}(B)$ . Similarly,  $A_0 \in \text{span}(A)$  which establishes (a).

To show (b), when  $\tilde{q} \geq q_0$ , from (a) we have  $B_0 \in \text{span}(B)$  and

$$\max_{A \in \mathcal{O}_{p,\tilde{p}}, B \in \mathcal{O}_{q,\tilde{q}}} E \|A^T(X - \mu)B\|_F^2 = \max_{A \in \mathcal{O}_{p,\tilde{p}}} \text{trace}(A^T E[A_0 U U^T A_0^T] A) + \tilde{p} \tilde{q} \sigma^2, \tag{A.2}$$

which is an eigenvalue-problem for the matrix  $E[A_0 U U^T A_0^T]$ . By diagonalizing  $E[U U^T] = \Gamma_U \Lambda_U \Gamma_U^T$ , then,  $E[A_0 U U^T A_0^T]$  has  $p_0$  non-zero eigenvalues  $\Lambda_U$  with the corresponding

eigenvectors  $A_0\Gamma_U$ . When  $\tilde{p} < p_0$ , the maximizer  $A$  consists of the first  $\tilde{p}$  columns of  $A_0\Gamma_U$  and, hence,  $\text{span}(A) \subsetneq \text{span}(A_0)$ .

(c) can be established in a similar way as (b).

To show (d), observe that

$$E\|A^T(X - \mu)B\|_F^2 = E\{\text{trace}(A^T A_0 U B_0^T B B^T B_0 U^T A_0^T A)\} + \tilde{p}\tilde{q}\sigma^2. \quad (\text{A.3})$$

To maximize (A.3) over  $A \in \mathcal{O}_{p,\tilde{p}}, B \in \mathcal{O}_{q,\tilde{q}}$  with  $\tilde{p} < p_0$  and  $\tilde{q} < q_0$ , the rank of  $A^T A_0$  and  $B^T B_0$  must be  $\tilde{p}$  and  $\tilde{q}$ , respectively, in order to attain the maximal value. This can happen only if  $\text{span}(A) \subsetneq \text{span}(A_0)$  and  $\text{span}(B) \subsetneq \text{span}(B_0)$ .  $\square$

**Proof of Proposition 2.6.** We will only provide a proof for (a), and (b) can be obtained in a similar way. If  $\tilde{q} \geq q_0$ , from Proposition 2.5 (a) we have  $\text{span}(B_0) \subseteq \text{span}(B)$ , which further implies that

$$E[(X - \mu)P_B(X - \mu)^T] = E[A_0 U U^T A_0^T] + E[\varepsilon P_B \varepsilon^T], \quad (\text{A.4})$$

$$E[(X - \mu)(X - \mu)^T] = E[A_0 U U^T A_0^T] + E[\varepsilon \varepsilon^T]. \quad (\text{A.5})$$

Note that  $E[\varepsilon P_B \varepsilon^T] = \tilde{q}\sigma^2 I_p$  and  $E[\varepsilon \varepsilon^T] = q\sigma^2 I_p$ . Hence,  $E[(X - \mu)(X - \mu)^T]$  and  $E[(X - \mu)P_B(X - \mu)^T]$  have the same leading  $p_0 \wedge \tilde{p}$  eigenvectors as  $E[A_0 U U^T A_0^T]$  has. Moreover, we have  $\lambda_i = d_i + \tilde{q}\sigma^2$  and  $\lambda_i^* = d_i + q\sigma^2$ , where  $d_i$  is the  $i^{\text{th}}$  eigenvalue of  $E[A_0 U U^T A_0^T]$ . Hence,  $\lambda_i^* - \lambda_i = (q - \tilde{q})\sigma^2$  for  $i = 1, \dots, \tilde{p}$ , which completes the proof.  $\square$

**Proof of Theorem 3.1.** Let  $H_{\tilde{p},\tilde{q}}(S_n) = (\text{vec}(\hat{A})^T, \text{vec}(\hat{B})^T)^T$  be the function maps  $S_n$  to its tensor principal components under  $(\tilde{p}, \tilde{q})$ , which gives  $H_{\tilde{p},\tilde{q}}(\Sigma) = (\text{vec}(A)^T, \text{vec}(B)^T)^T$  and  $D_{H_{\tilde{p},\tilde{q}}} = \frac{\partial H_{\tilde{p},\tilde{q}}(\Sigma)}{\partial \text{vec}(\Sigma)}$ . Note that  $\hat{\Phi}(p, q) = \text{vec}(I_m)^T \text{vec}(S_n)$  and  $\Phi(p, q) = \text{vec}(I_m)^T \text{vec}(\Sigma)$ . From the weak convergence  $\sqrt{n}(S_n - \Sigma) \xrightarrow{d} N$  and an application of the delta method, we have, for  $(\tilde{p}, \tilde{q}) \leq (p_0, q_0)$ ,

$$\begin{aligned} \sqrt{n} \left( \hat{\Phi}(p, q) - \Phi(p, q) \right) &\xrightarrow{d} \text{vec}(I_m)^T \text{vec}(N), \\ \sqrt{n} \left( \hat{\Phi}(\tilde{p}, \tilde{q}) - \Phi(\tilde{p}, \tilde{q}) \right) &\xrightarrow{d} D_{\Phi(\tilde{p}, \tilde{q})} \text{vec}(N), \\ \sqrt{n} (H_{\tilde{p},\tilde{q}}(S_n) - H_{\tilde{p},\tilde{q}}(\Sigma)) &\xrightarrow{d} D_{H_{\tilde{p},\tilde{q}}} \text{vec}(N). \end{aligned}$$

The explicit forms of  $D_{\Phi(\tilde{p}, \tilde{q})}$  and elements in  $D_{H_{p_0, q_0}}$  are provided in Lemma 3.2.  $\square$

**Proof of Lemma 3.2.** For a given pair  $(\tilde{p}, \tilde{q})$  with  $1 \leq \tilde{p} \leq p$  and  $1 \leq \tilde{q} \leq q$ , we have, from (6) and (7), that  $A$  and  $B$  satisfy the following system of stationary equations

$$\begin{aligned} \left( \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p) \right) a_i &= \lambda_i a_i, \quad i = 1, \dots, \tilde{p}, \\ \left( \sum_{i=1}^{\tilde{p}} (I_q \otimes a_i)^T \Sigma (I_q \otimes a_i) \right) b_j &= \xi_j b_j, \quad j = 1, \dots, \tilde{q}, \end{aligned}$$

where  $a_i, b_j, \lambda_i, \xi_j$  depend on  $(\Sigma, \tilde{p}, \tilde{q})$ . The indices  $i, j$  in the above system of equations can go beyond  $\tilde{p}$  and  $\tilde{q}$  and up to  $p$  and  $q$ . But those  $a_i$  and  $b_j$  with  $i > \tilde{p}$  and  $j > \tilde{q}$  will not be included in the solution pair  $(A, B)$ . Note that we have the following identity, which is due to the definition of  $\Phi$  and the stationary equations:

$$\Phi(\tilde{p}, \tilde{q}) = \sum_{i=1}^{\tilde{p}} \lambda_i(\Sigma, \tilde{p}, \tilde{q}) = \sum_{j=1}^{\tilde{q}} \xi_j(\Sigma, \tilde{p}, \tilde{q}). \quad (\text{A.6})$$

We will use the perturbation method (Sibson, Lemma 2.1, 1979; Fine, 1987) to derive the derivatives  $D_{\Phi(\tilde{p}, \tilde{q})}$ ,  $\frac{\partial a_i}{\partial \text{vec}(\Sigma)}$  and  $\frac{\partial b_j}{\partial \text{vec}(\Sigma)}$ . Suppose that  $\Sigma$  is perturbed to  $\Sigma_\epsilon = \Sigma + \epsilon \dot{\Sigma}$ . Denote the corresponding system of stationary equations with  $\Sigma_\epsilon$  by

$$\left( \sum_{j=1}^{\tilde{q}} (b_{j,\epsilon} \otimes I_p)^T \Sigma_\epsilon (b_{j,\epsilon} \otimes I_p) \right) a_{i,\epsilon} = \lambda_{i,\epsilon} a_{i,\epsilon}, \quad i = 1, \dots, \tilde{p}, \quad (\text{A.7})$$

$$\left( \sum_{i=1}^{\tilde{p}} (I_q \otimes a_{i,\epsilon})^T \Sigma_\epsilon (I_q \otimes a_{i,\epsilon}) \right) b_{j,\epsilon} = \xi_j b_{j,\epsilon}, \quad j = 1, \dots, \tilde{q}. \quad (\text{A.8})$$

Let their first order expansions be denoted by

$$\begin{aligned} \lambda_{i,\epsilon} &= \lambda_i + \epsilon \dot{\lambda}_i + o(\epsilon), & a_{i,\epsilon} &= a_i + \epsilon \dot{a}_i + o(\epsilon), \\ \xi_{j,\epsilon} &= \xi_j + \epsilon \dot{\xi}_j + o(\epsilon), & b_{j,\epsilon} &= b_j + \epsilon \dot{b}_j + o(\epsilon). \end{aligned}$$

Following the same arguments as in Lemma 2.1 of Sibson (1979) and by equating the terms involving  $\epsilon$  in (A.7) we have, for  $i = 1, \dots, \tilde{p}$ ,

$$\dot{\lambda}_i = a_i^T \dot{\Sigma}_B a_i, \quad (\text{A.9})$$

$$\dot{a}_i = \left\{ \lambda_i I_p - \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p) \right\}^+ \dot{\Sigma}_B a_i, \quad (\text{A.10})$$

where

$$\begin{aligned}
\dot{\Sigma}_B &= \sum_{j=1}^{\tilde{q}} \left( (\dot{b}_j \otimes I_p)^T \Sigma (b_j \otimes I_p) + (b_j \otimes I_p)^T \Sigma (\dot{b}_j \otimes I_p) \right) + \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \dot{\Sigma} (b_j \otimes I_p) \\
&= E[X(\dot{B}B^T + B\dot{B}^T)X^T] + \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \dot{\Sigma} (b_j \otimes I_p). \tag{A.11}
\end{aligned}$$

Since  $B_\epsilon^T B_\epsilon = I_{\tilde{q}}$ , then  $\dot{B} = [\dot{b}_1, \dots, \dot{b}_{\tilde{q}}]$  must satisfy  $\dot{B}^T B + B^T \dot{B} = 0$ .

(a) For  $(\tilde{p}, \tilde{q}) \leq (p_0, q_0)$ , the first term of  $\sum_{i=1}^{\tilde{p}} a_i^T \dot{\Sigma}_B a_i$  can be expressed as

$$\sum_{i=1}^{\tilde{p}} a_i^T E[X(\dot{B}B^T + B\dot{B}^T)X^T] a_i = \sum_{j=1}^{\tilde{q}} \left( \dot{b}_j^T E[X^T P_A X] b_j + b_j^T E[X^T P_A X] \dot{b}_j \right)$$

which vanishes by noting that  $b_j$  is an eigenvector of  $E[X P_A X^T]$  and  $b_j^T \dot{b}_j = \dot{b}_j^T b_j = 0$ .

This concludes that  $\sum_{i=1}^{\tilde{p}} \dot{\lambda}_i = \sum_{i=1}^{\tilde{p}} a_i^T \left( \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \dot{\Sigma} (b_j \otimes I_p) \right) a_i$  and, hence,

$$D_{\Phi(\tilde{p}, \tilde{q})} = \sum_{i=1}^{\tilde{p}} \sum_{j=1}^{\tilde{q}} (b_j \otimes a_i \otimes b_j \otimes a_i)^T = \text{vec}(P_{B \otimes A})^T. \tag{A.12}$$

(b) Assume now  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ . To derive the form of  $\frac{\partial a_i}{\partial \text{vec}(\Sigma)}$ , we are going to show that the first term of  $\dot{\Sigma}_B$  is zero and conclude  $\dot{\Sigma}_B = \sum_{j=1}^{q_0} (b_j \otimes I_p)^T \dot{\Sigma} (b_j \otimes I_p)$ . This together with (A.10) gives

$$\begin{aligned}
\frac{\partial a_i}{\partial \text{vec}(\Sigma)} &= \sum_{j=1}^{q_0} (b_j \otimes a_i \otimes b_j \otimes \{\lambda_i I_p - E[X P_B X^T]\}^+)^T \\
&= \{a_i \otimes \text{vec}(P_{B_0}) \otimes (\lambda_i I_p - E[X P_{B_0} X^T])^+\}^T (K_{p,q} \otimes I_{pq}). \tag{A.13}
\end{aligned}$$

as desired, where the second equality follows from Proposition 2.5 that  $\text{span}(B) = \text{span}(B_0)$  when  $\tilde{q} = q_0$ . To complete the proof, first note that Proposition 2.5 ensures the existence of a nonsingular matrix  $\eta$  such that  $B_0 = B\eta$ . From  $X = A_0 U B_0^T + \varepsilon$  (remember  $\mu = 0$ ) and the independent structure of  $U$  and  $\varepsilon$ , we can represent the first term of  $\dot{\Sigma}_B$  as

$$\begin{aligned}
E[X(\dot{B}B^T + B\dot{B}^T)X^T] &= E[A_0 U B_0^T (\dot{B}B^T + B\dot{B}^T) B_0 U^T A_0^T] + \sigma^2 \text{trace}(B^T \dot{B} + \dot{B}^T B) I_p \\
&= E[A_0 U \eta^T (B^T \dot{B} + \dot{B}^T B) \eta U^T A_0^T] + \sigma^2 \text{trace}(B^T \dot{B} + \dot{B}^T B) I_p.
\end{aligned}$$

The proof is completed by noting that  $B^T \dot{B} + \dot{B}^T B = 0$ . The case of  $\frac{\partial b_j}{\partial \text{vec}(\Sigma)}$  can be established in a similar way.  $\square$



**Proof of Corollary 3.3.** Consider the function  $F(A, B) = P_{B \otimes A}$  with the corresponding differential  $D_{F(A, B)}$ . From Theorem 3.1 (b) and delta method, we have

$$\begin{aligned} \sqrt{n} \text{vec}(P_{\hat{B} \otimes \hat{A}} - P_{B \otimes A}) &= \sqrt{n} \text{vec}(F(\hat{A}, \hat{B}) - F(A, B)) \\ &= D_{F(A, B)} \sqrt{n} \left( \begin{bmatrix} \text{vec}(\hat{A}) \\ \text{vec}(\hat{B}) \end{bmatrix} - \begin{bmatrix} \text{vec}(A) \\ \text{vec}(B) \end{bmatrix} \right) + o\left(\frac{1}{\sqrt{n}}\right) \\ &\xrightarrow{d} D_{P_{B \otimes A}} \text{vec}(N), \end{aligned}$$

where  $D_{P_{B \otimes A}} = D_{F(A, B)} D_{H_{\tilde{p}, \tilde{q}}}$ . When  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ , the expression of  $D_{P_{B \otimes A}}$  is obtained by a direct calculation together with Lemma 3.2 (b) and Theorem 2.5 (a).  $\square$

**Proof of Corollary 3.4.** Consider the function  $F(x, y) = x/y$  with the corresponding differential  $D_{F(x, y)} = (y^{-1}, -xy^{-2})^T$ . From Theorem 3.1 (a) and delta method, we have

$$\begin{aligned} \sqrt{n}(\hat{\rho}(\tilde{p}, \tilde{q}) - \rho(\tilde{p}, \tilde{q})) &= \sqrt{n}(F(\hat{\Phi}(\tilde{p}, \tilde{q}), \hat{\Phi}(p, q)) - F(\Phi(\tilde{p}, \tilde{q}), \Phi(p, q))) \\ &= D_{F(\Phi(\tilde{p}, \tilde{q}), \Phi(p, q))} \sqrt{n} \left( \begin{bmatrix} \hat{\Phi}(\tilde{p}, \tilde{q}) \\ \hat{\Phi}(p, q) \end{bmatrix} - \begin{bmatrix} \Phi(\tilde{p}, \tilde{q}) \\ \Phi(p, q) \end{bmatrix} \right) + o\left(\frac{1}{\sqrt{n}}\right) \\ &\xrightarrow{d} D_{F(\Phi(\tilde{p}, \tilde{q}), \Phi(p, q))} \begin{bmatrix} D_{\Phi(\tilde{p}, \tilde{q})} \\ \text{vec}(I_m)^T \end{bmatrix} \text{vec}(N). \end{aligned}$$

A direct calculation gives the expression of the asymptotic variance  $\sigma_{\rho(\tilde{p}, \tilde{q})}^2$ .  $\square$

**Proof of Theorem 3.5.** Under  $H_0$ , we have from Corollary 3.4 that, for  $n$  large enough,

$$P \left( \frac{\sqrt{n}(\hat{\rho}(\tilde{p}, \tilde{q}) - \rho_0)}{\sigma_{\rho(\tilde{p}, \tilde{q})}} > z_\alpha \right) \leq \alpha.$$

The consistency of  $\hat{\sigma}_{\rho(\tilde{p}, \tilde{q})}^2$  and Slutsky's theorem complete the proof.  $\square$

**Proof of Theorem 3.6.** Since  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$ , we have  $A = A^*$  and  $B = B^*$  from Theorem 2.6, and  $\text{span}(A) = \text{span}(A_0)$  and  $\text{span}(B) = \text{span}(B_0)$  from Theorem 2.5 (a). Let  $\{a_i : i > p_0\}$  and  $\{b_j : j > q_0\}$  be orthogonal bases of  $\text{span}(Q_{A_0})$  and  $\text{span}(Q_{B_0})$ ,  $W_{B, q'} = \sum_{j=1}^{q'} (b_j \otimes I_p \otimes b_j \otimes I_p)^T$ ,  $q' = 1, \dots, q$  and  $W_{A, p'} = \sum_{i=1}^{p'} (I_q \otimes a_i \otimes I_q \otimes a_i)^T$ ,  $p' = 1, \dots, p$ . Also define  $M_A = [(a_1 \otimes M_{A1}), \dots, (a_{p_0} \otimes M_{A p_0})]^T$  and  $M_B = [(b_1 \otimes M_{B1}), \dots, (b_{q_0} \otimes M_{B q_0})]^T$ , where  $M_{Ai} = \{\lambda_i I_p - E[X P_{B_0} X^T]\}^+$  and  $M_{Bj} = \{\xi_j I_q -$

$E[X^T P_{A_0} X]\}^+$ . By using these notations and from Theorem 3.1 (b), we have the limiting distribution of MPCA

$$\sqrt{n}(\text{vec}(\hat{A}, \hat{B}) - \text{vec}(A, B)) \xrightarrow{d} M_0 W_0 \text{vec}(N), \quad (\text{A.14})$$

where  $M_0 = \begin{bmatrix} M_A & 0 \\ 0 & M_B \end{bmatrix}$ ,  $W_0 = \begin{bmatrix} W_{B,q_0} \\ W_{A,p_0} \end{bmatrix}$ . By Lemma A.1 below, the limiting distribution of (2D)<sup>2</sup>PCA is derived to be

$$\sqrt{n}(\text{vec}(\hat{A}^*, \hat{B}^*) - \text{vec}(A^*, B^*)) \xrightarrow{d} M_0(W_0 + W_{0+})\text{vec}(N), \quad (\text{A.15})$$

where  $W_{0+} = \begin{bmatrix} W_{B,q_0+} \\ W_{A,p_0+} \end{bmatrix}$ ,  $W_{A,p_0+} = W_{A,p} - W_{A,p_0}$ , and  $W_{B,q_0+} = W_{B,q} - W_{B,q_0}$ .

Note that  $\hat{A} = A^*$  and  $\hat{B} = B^*$ . To complete the proof, by an application of delta method, it thus suffices to show

$$\text{aCov}(\text{vec}(\hat{A}^*, \hat{B}^*)) - \text{aCov}(\text{vec}(\hat{A}, \hat{B})) \geq 0. \quad (\text{A.16})$$

From (A.14)-(A.15) we are left to show

$$M_0(W_0 \Sigma_N W_{0+}^T + W_{0+} \Sigma_N W_0^T + W_{0+} \Sigma_N W_{0+}^T) M_0^T \geq 0, \quad (\text{A.17})$$

where  $\Sigma_N = \text{Cov}(\text{vec}(N)) = (I_{m^2} + K_{m,m})(\Sigma \otimes \Sigma)$  under normality of  $\text{vec}(X)$ . We are going to show  $M_0 W_0 \Sigma_N W_{0+}^T M_0^T = 0$ . This together with the fact  $M_0 W_{0+} \Sigma_N W_{0+}^T M_0^T \geq 0$  then establishes the desired result. Observe that

$$\begin{aligned} & M_0 W_0 \Sigma_N W_{0+}^T M_0^T \\ &= \begin{bmatrix} M_A W_{B,q_0} \Sigma_N W_{B,q_0+}^T M_A^T & H_A W_{B,q_0} \Sigma_N W_{A,p_0+}^T M_B^T \\ H_B W_{A,p_0} \Sigma_N W_{B,q_0+}^T M_A^T & H_B W_{A,p_0} \Sigma_N W_{A,p_0+}^T M_B^T \end{bmatrix}. \end{aligned}$$

From model (1),  $\Sigma = (B_0 \otimes A_0)(T + \sigma^2 I_{m_0})(B_0 \otimes A_0)^T + \sigma^2 Q_{B_0 \otimes A_0}$ , where  $Q_{B_0 \otimes A_0} = Q_{B_0} \otimes P_{A_0} + P_{B_0} \otimes Q_{A_0} + Q_{B_0} \otimes Q_{A_0}$ . This implies  $W_{B,q_0} \Sigma_N W_{B,q_0+}^T = 0$  and  $W_{A,p_0} \Sigma_N W_{A,p_0+}^T = 0$  and, hence, the diagonal elements of the above matrix vanish. For the off-diagonal elements, the same reasoning can be used to deduce that  $(M_A W_{B,q_0}) \Sigma_N W_{A,p_0+}^T = 0$  and  $(M_B W_{A,p_0}) \Sigma_N W_{B,q_0+}^T = 0$ , which establishes (A.16). A direct calculation further gives

$$M_0 W_{0+} \Sigma_N W_{0+}^T M_0^T = \sigma^4 \begin{bmatrix} (q - q_0) M_A (I_{p^2} + K_{p,p}) M_A^T & 0 \\ 0 & (p - p_0) M_B (I_{q^2} + K_{q,q}) M_B^T \end{bmatrix},$$

which equals a zero matrix if and only if  $(p_0, q_0) = (p, q)$ .  $\square$

**Lemma A.1.** *Assume model (1) and assume that the leading eigenvalues  $\{\lambda_i^* : i = 1, \dots, p_0\}$  and  $\{\xi_j^* : j = 1, \dots, q_0\}$  of  $(2D)^2PCA$  are simple roots. Then, the differentials of  $(2D)^2PCA$  components with respect to  $\Sigma$  under  $(\tilde{p}, \tilde{q}) = (p_0, q_0)$  are given by*

$$\frac{\partial \text{vec}(A^*)}{\partial \text{vec}(\Sigma)} = M_A W_{B,q} \quad \text{and} \quad \frac{\partial \text{vec}(B^*)}{\partial \text{vec}(\Sigma)} = M_B W_{A,p}. \quad (\text{A.18})$$

*Proof.* We only derive the differential of  $A^*$ , where the case of  $B^*$  is similarly obtained. Remember that  $(2D)^2PCA$  components  $A^*$  are leading eigenvectors of  $K_{A^*} = E[XX^T]$  with eigenvalues  $\lambda_i^*$ . A standard argument (Sibson, 1979) then gives

$$\begin{aligned} \frac{\partial \text{vec}(a_i^*)}{\partial \text{vec}(K_{A^*})} &= a_i^{*T} \otimes \{\lambda_i^* I_p - K_{A^*}\}^+ \\ &= a_i^T \otimes M_{Ai}, \end{aligned} \quad (\text{A.19})$$

where the second equality follows from Theorem 2.6 with  $M_{Ai}$  being defined in Theorem 3.6. Turning to the differential of  $K_{A^*}$  with respect to  $\Sigma$ . It is always true that

$$K_{A^*} = \sum_{j=1}^q (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p),$$

where  $\{b_j : j > q_0\}$  are defined in the beginning of Theorem 3.6. Thus, we have

$$\frac{\partial \text{vec}(K_{A^*})}{\partial \text{vec}(\Sigma)} = W_{B,q}. \quad (\text{A.20})$$

From (A.19)-(A.20) and the chain rule, the proof is completed.  $\square$



Figure 1: 40 test faces randomly drawn from the test set.



Figure 2: Reconstructed faces using  $28 \times 28$  trained MPCA components.



Figure 3: Reconstructed faces using 784 trained conventional PCA components.

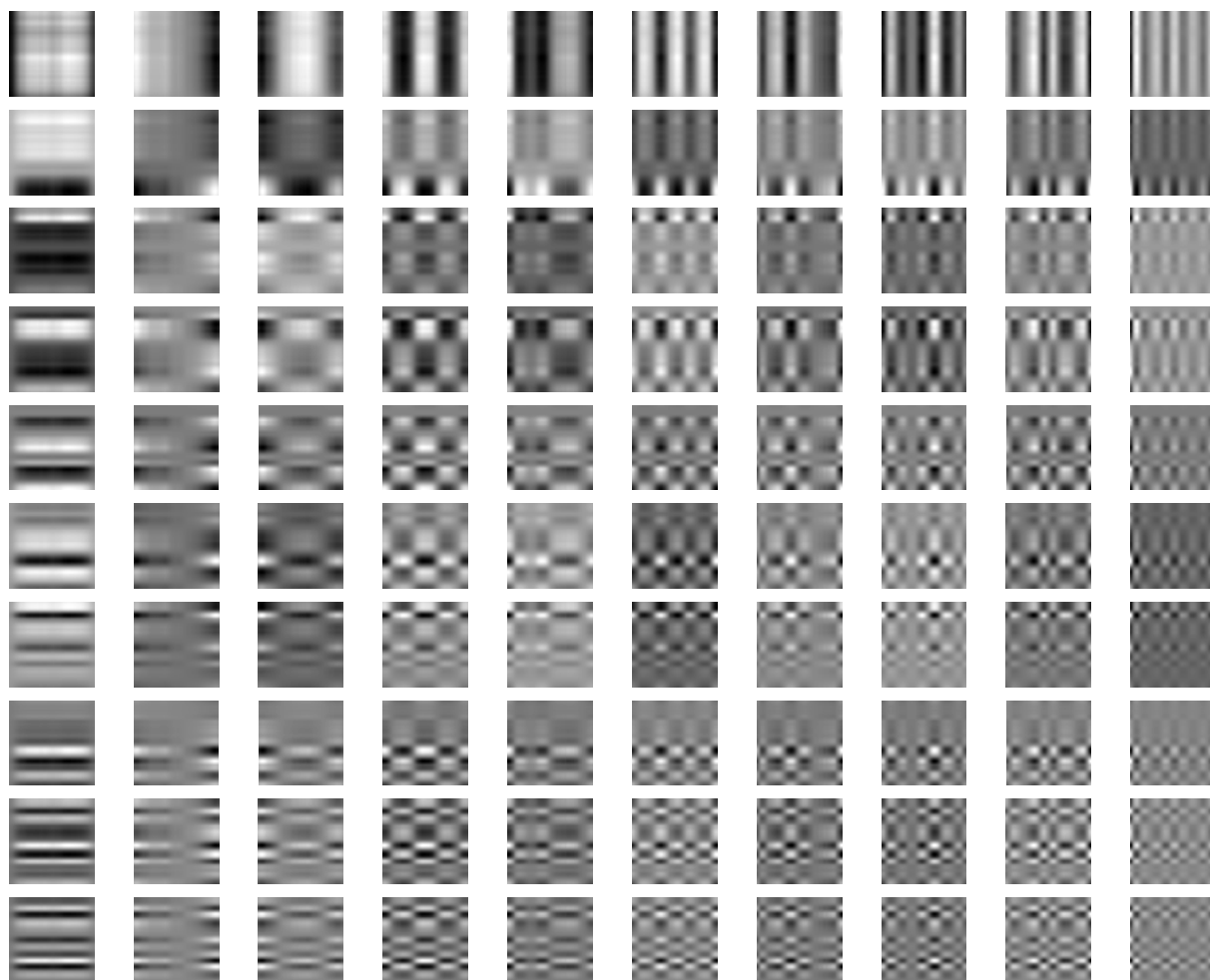


Figure 4: Leading 100 MPCA basis images.



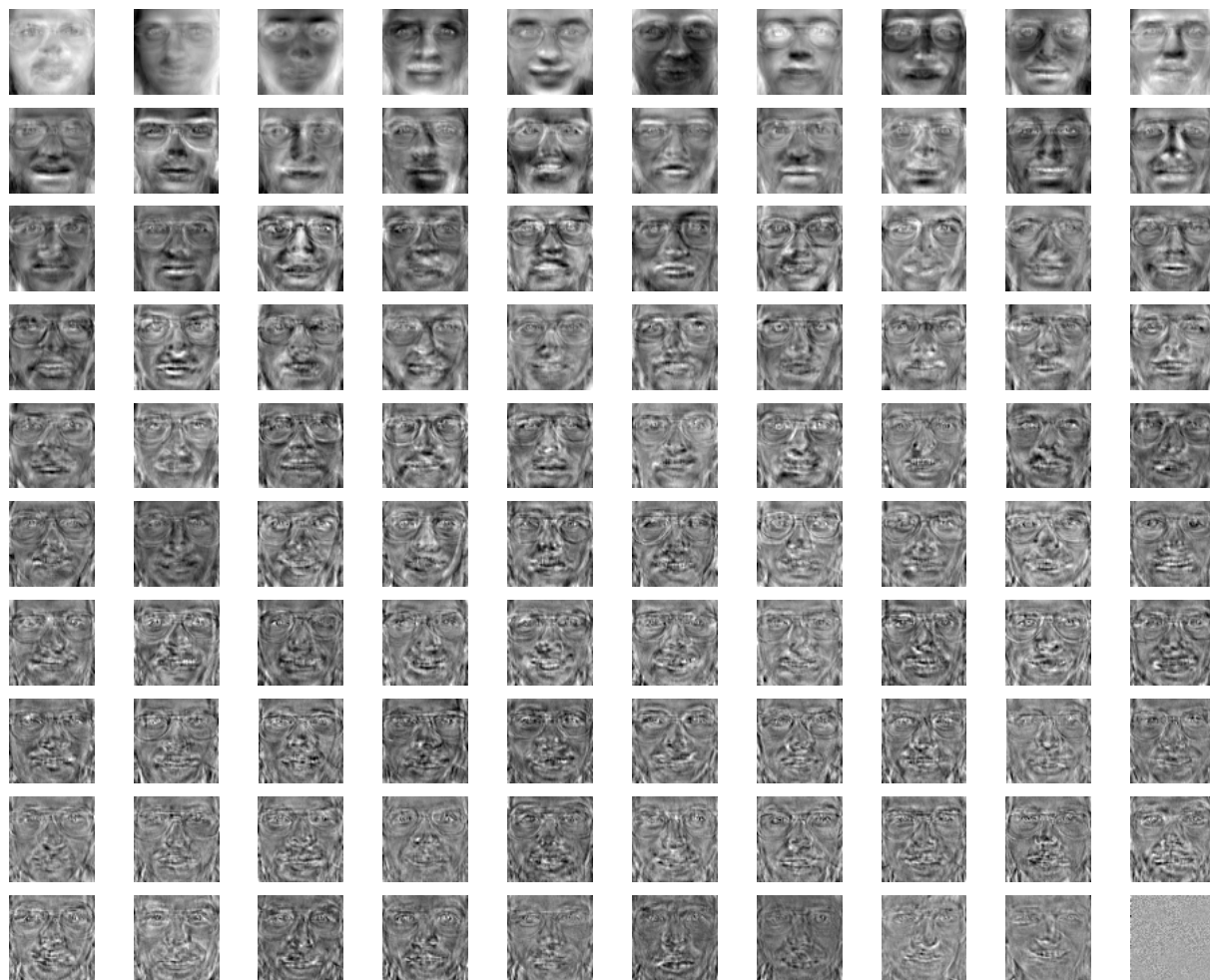


Figure 5: Leading 100 conventional PCA basis images.



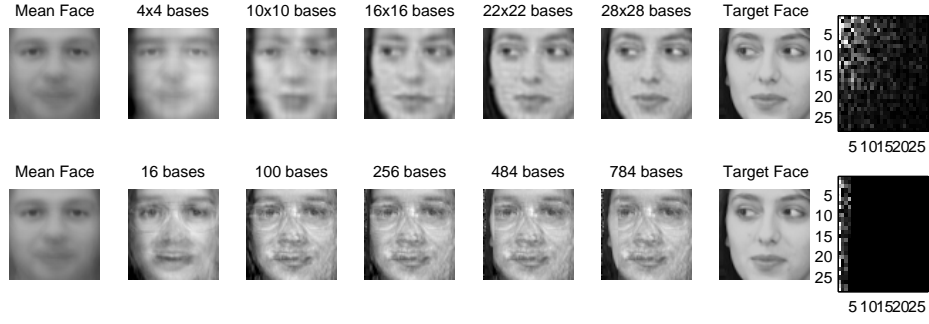


Figure 6: The reconstructed images for the test face by adding more basis elements are compared for MPCA (top row) and conventional PCA (bottom row). Both mean face and the target face are put in this figure as references. Plots in the right-most column are projection coefficients (in absolute value) onto PCA subspace.

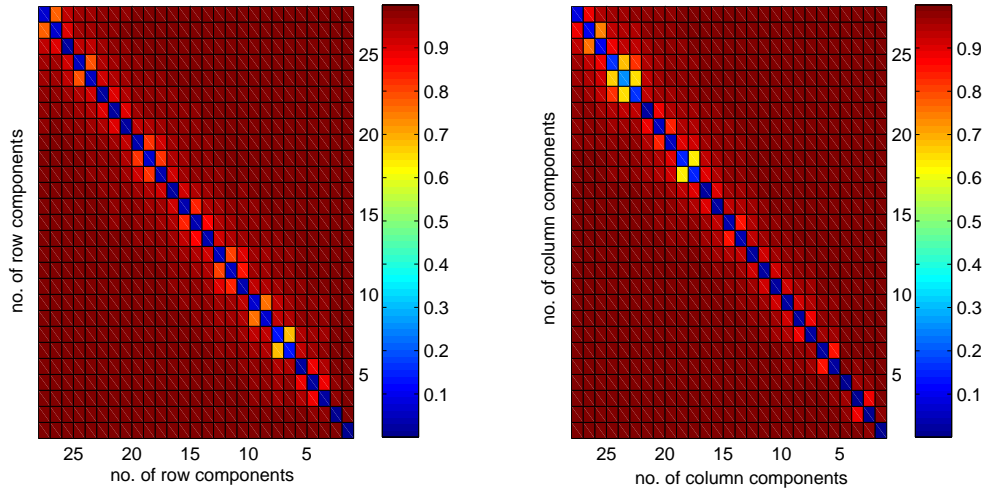


Figure 7: The variability pattern plots for MPCA for  $1 \leq \tilde{p} \leq 28$  and  $1 \leq \tilde{q} \leq 28$ .